Recommending What Video to Watch Next: Candidate Generation, A Multitask Ranking System

Kavithaa Suresh Kumar
ks64@illinois.edu

1. Introduction

In this technical review, the recommendation system algorithm used by YouTube application is discussed. YouTube's recommendation system recommends a personalised list of videos for its billion users. The recommendation algorithm narrows down a corpus of billions of videos to a ranked list of few hundred relevant videos to the user. Recommending a list of videos that are personalised to each user's interest is challenging due to the following reasons:

- to optimise engagement objectives (e.g., clicks)
- to optimise satisfaction objectives (e.g., likes)
- to reduce selection bias

To overcome these challenges YouTube uses two deep neural networks in its algorithm that are described below.

2. System overview

The recommendation system consists of two neural networks: Candidate generation and Multitask ranking system. First is the candidate generation, that takes inputs like entire corpus of YouTube videos along with user's watch history, search history, demographic. High precision is achieved in this stage by recommending a small subset of relevant videos (few hundreds) to the users. Second is Multitask ranking system, it uses the output of previous system as its input to calculate a score for each video. These generated scores are used to rank the list of videos in order for the system to achieve high recall. This generated ranked list is used to recommend the videos to the user.

3. Candidate generation

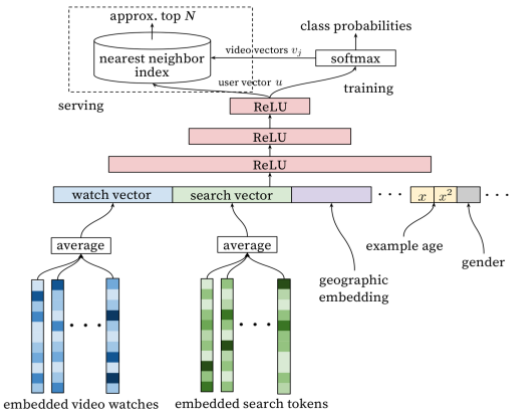

Figure 1 [1]: Candidate generation Architecture

Candidate generation network narrows down YouTube's entire corpus of videos to a list of few hundreds of personalised videos. This network uses user's watch history, watch time, gender, demographics etc. as inputs. Another input used by this system is a list of videos watched by similar users and this is generated by collaborative filtering. The similarity between users is measured with ID's of watched videos by similar users.
Figure 1 shows the general architecture of Candidate generation neural network. Candidate generation has multiple stage. First stage takes item embeddings of videos history and search tokens as input. These embeddings are high dimensional. As the number of watches or search token varies for each user, an average of these embeddings is calculated. This calculated average is used as fixed size input for the second stage i.e., combiner stage. In this combiner stage concatenation of embeddings of different features takes place. The features include knowledge about the users like location, gender, age of the videos etc. This concatenation is helpful to handle unique cases. It uses other embeddings for prediction to handle unique cases like user with no watch history. Third stage is a tower architecture made of multiple layers of fully connected Rectified linear unit (ReLU). This layer is made up of flops to perform computation. The ReLU layer generates a user vector as the output. Fourth stage is the Softmax function that predicts multinomial distribution i.e., likelihood of the video being watched by the user. The Softmax outputs the video vector based on the prediction. During serving time video vector and user vector is combined along with Nearest neighbor to choose top N videos to the user.

4. Multitask ranking system

The multitask ranking system takes the output of candidate generation as input and outputs a ranked list of videos to the user. The main focus is to optimise the Engagement objectives and Satisfaction objectives. Engagement objectives is measured with user clicks, watch time. Satisfaction objectives is measured with likes, shares, comments, dismissal, etc. Objectives are classified as below

- binary classification: click or not, like or dismissal
- regression: watch time, rating given etc.

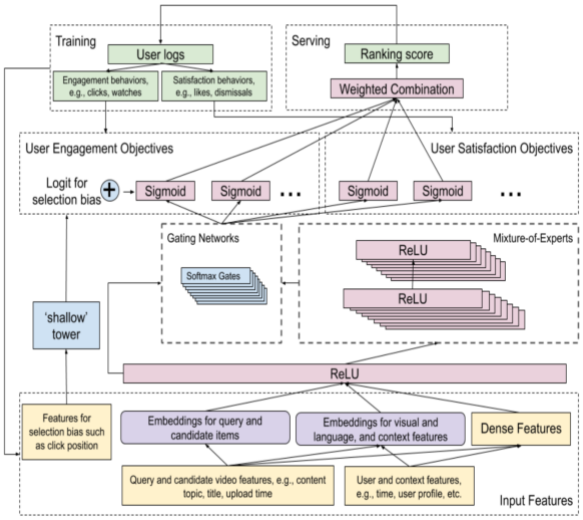

Figure 2[2]: Multitask ranking system Architecture

The Multitask ranking system architecture is shown in figure 2. The input layer consists of embeddings of content, title of current videos along with user watch time, user profile etc. This input is fed to the Multi gate Mixture of experts (MMoE)layer. MMoE consists of two components: Mixture of experts and Gating networks. The Mixture of experts component has multiple Expert layers. The input for this component is from the previous input layer. Each expert layer is used for different objective and it learns to become an expert for a specific input. Another component is Gating network made of many softmax functions. The input for these softmax is fed from the expert layers. Different experts are needed for different objectives. This is achieved by the softmax function which determines the expert layers that are important for different objectives. The output of the MMoE layer is fed into two components named User Engagement Objective and User Satisfaction Objectives. These components are represented by sigmoid function. The output of each sigmoid function is a probability the video will be watched. At the serving time, these predicted values from sigmoid functions are used to calculate a score by either averaging the value or other functions. Once the scores are calculated for all videos from the candidate generation, the list is sorted in increasing order. Based on the sorted list, the videos are suggested to users in the ranked manner. The video interactions of the users are logged and used as training data in the training phase.

The biggest challenge with the above architecture is that only first few videos are watched by the user even though a video at the very bottom of the ranked list might be more interesting to the user. This is called as position bias. To remove the position bias, a shallow tower layer is introduced. During training phase, the position of the video is inputted to this shallow tower layer along with the device type. The logit is calculated and fed as an input to engagement objectives to train the network to remove position bias.

5. Conclusion

In this technical review we have discussed the architecture of two deep neural networks used in YouTube's recommendations system. The first neural network-candidate generation is used to reduce the entire corpus of videos to few hundred videos. The second neural network- Multitask ranking system consists of two important components MMoE and shallow tower to achieve the following objectives:

- to optimize User Engagement objectives
- to optimize User satisfaction objectives
- to remove position bias

By using the MMoE model and increasing the number of experts leads to an improvement in the performance of the system.

6. References

1.[1] Paul Covington, Jay Adams, Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations ACM ISBN 978-1-4503-4035-9/16/09.

2.[2] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, Ed Chi. 2019. Recommending What Video to Watch Next: A Multitask Ranking System. ACM ISBN 978-1-4503-6243-6/19/09